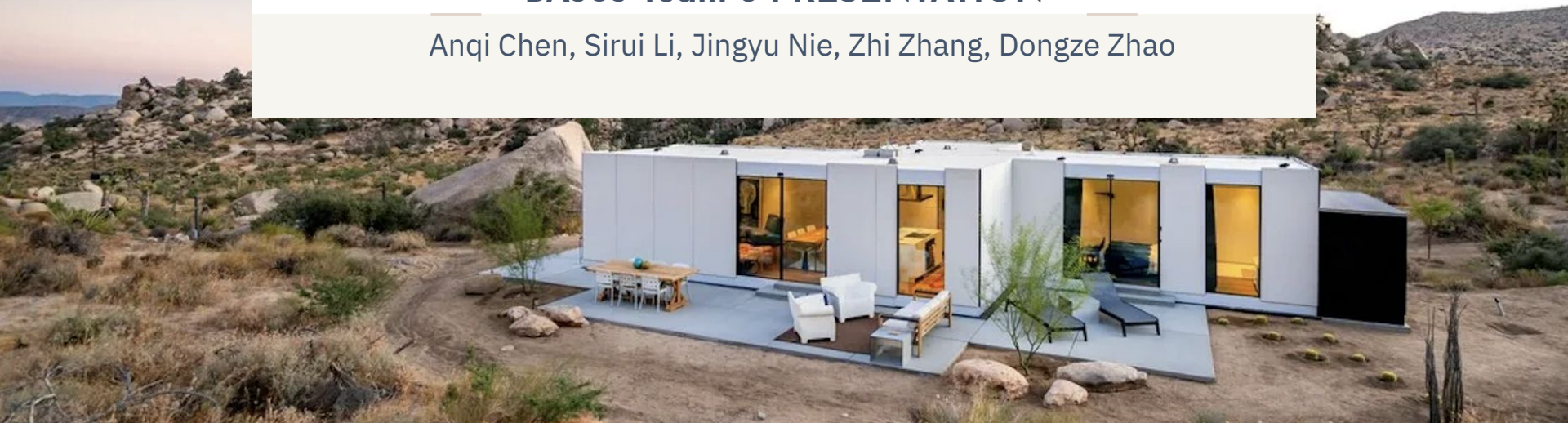
Europe Airbnb Price Prediction

BA305 Team 6 PRESENTATION

Anqi Chen, Sirui Li, Jingyu Nie, Zhi Zhang, Dongze Zhao



Agenda

1. Data Preprocessing
2. Data Visualization
3. Building Predictive Models
4. Model Evaluation

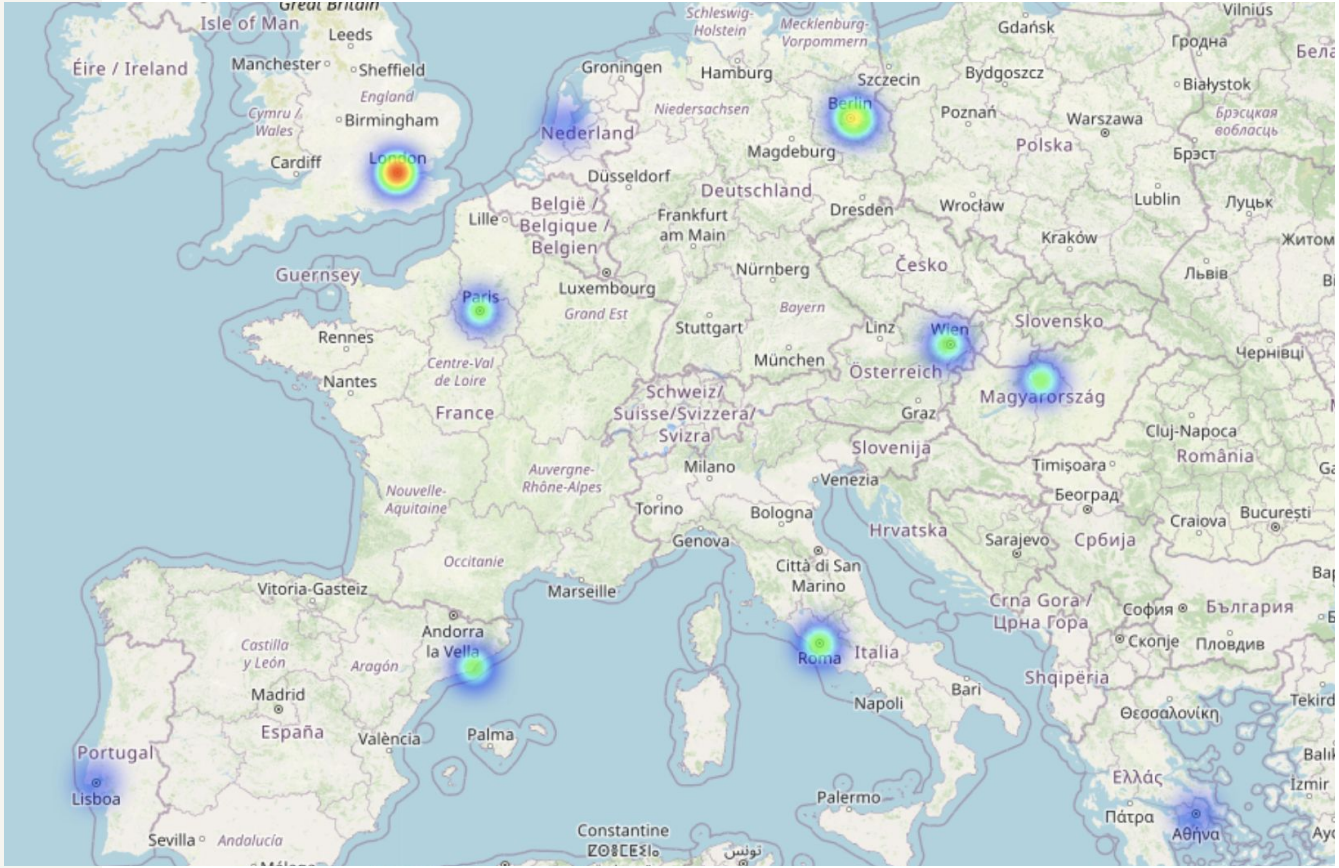
Mission Statement

Our mission is to offer practical insights into Airbnb pricing in Europe through our analytical models. By analyzing current market trends and price levels, We aim to

1. Assist homeowners in optimizing their pricing strategies,
2. Help potential real estate investors in making considerate investment decisions that maximize profits, and
3. Ultimately contribute to the growth and success of the European Airbnb market

Amsterdam
Athens
Barcelona
Berlin
Budapest
Lisbon
London
Paris
Rome
Vienna

10 Major
European Cities



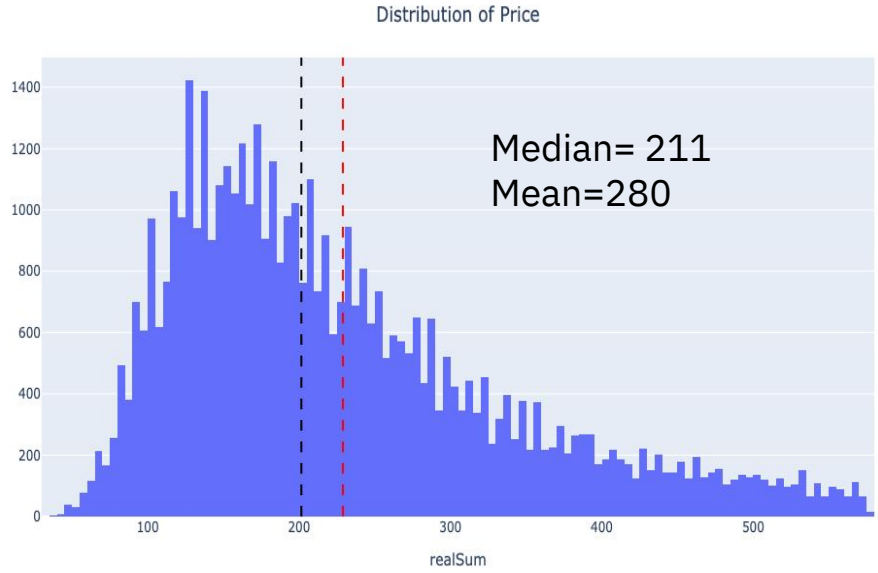
Factors that influence Airbnb Prices in European Cities

Column_Name	Description	Type	Sample	Decision	Reasons
realSum	Listing Price	Numeric	296.16	N/A	N/A
room_type	Private, Shared, & Entire Home	Categorical	Private Room	Converted to Dummy Variables	Make Interpretable for Analysis
room_shared	The Room is Shared?	Boolean	False	Dropped	Repetitive
room_private	The Room is Private?	Boolean	True	Dropped	Repetitive
person_capacity	-	Boolean	2.0	N/A	N/A
host_is_superhost	The Host is a Superhost?	Boolean	True	Converted to Numeric Value	Make Interpretable for Analysis
multi	For Multiple Rooms?	Boolean	0	N/A	N/A
biz	For Business Purposes?	Boolean	0	N/A	N/A
cleanliness_rating	-	Numeric	10.0	N/A	N/A
bedrooms	No. of Bedrooms	Numeric	1	N/A	N/A

Factors that influence Airbnb Prices in European Cities (Cont.)

Column_Name	Description	Type	Sample	Decision	Reasons
guest_satisfaction_overall	Satisfaction Rating	Numeric	97.0	N/A	N/A
dist	Distance from the City Center	Numeric	0.70	N/A	N/A
metro_dist	Distance from the Metro Station	Numeric	0.19	N/A	N/A
attr_index	Attraction Index	Numeric	518.48	Dropped	Redundant
attr_index_norm	Normalised Attraction Index	(0-100)	25.24	N/A	N/A
rest_index	Restaurant Index	Numeric	1218.66	Dropped	Redundant
rest_index_norm	Normalised Restaurant Index	(0-100)	71.61	N/A	N/A
lng	Longitude	Numeric	2.35	Dropped	Irrelevant
lat	Latitude	Numeric	48.86	Dropped	Irrelevant

Data Processing - identify outliers using the interquartile range method



Original dataframe shape: (51707, 22)

Filtered dataframe shape: (48045, 22)

Both distributions are **skewed to the right**

Data Cleaning - Before VS. After

Raw Data	
No. of Datasets	20 (10 Cities: Weekdays & Weekends)
No. of Variables	19
Records	51707
Missing Values / Duplicate Values	0

Processed Data	
No. of Datasets	Integrated into 1
No. of Variables	27
Records	48045
Missing Values / Duplicate Values	0
Variables Dropped (7)	room_type, room_shared, room_private, attr_index, rest_index, lng, lat
Convert to Dummy Variables (15)	3 room_type dummies, Weekend_dum_False, Weekend_dum_True, 10 City dummies

02

Data Visualization

Data Visualization - Prices Comparison

Finding 1 - Compare Average Prices in Different Cities

- The average price in Amsterdam is the highest at €370, while in Athens it is the lowest at €150.

Finding 2 - Compare the price distribution of each city

- Amsterdam, London, and Paris have the relatively high median price and wide interquartile ranges, indicating that there is a lot of variation in prices in these cities.

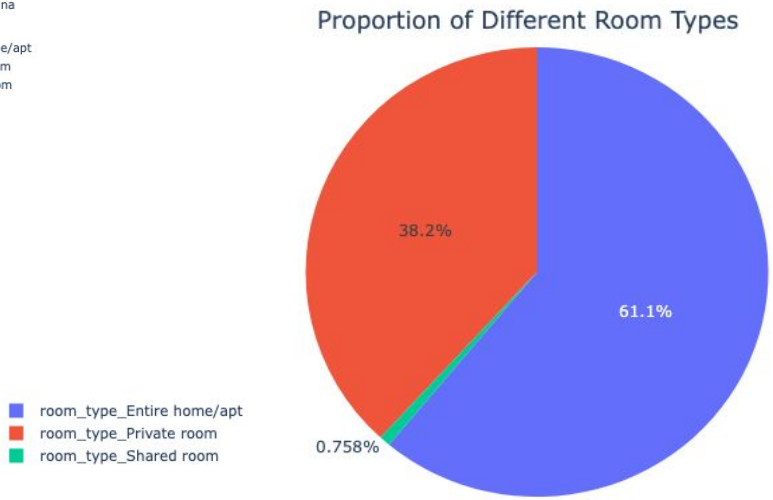
Data Visualization - Room Type Comparison

Finding 3 - Compare average price of different room types in each city

- The average price for a shared room is generally much lower than the other two room types in all cities.

Finding 4 - Compare the proportion of different room types

- Most of the people prefer entire home/apt.
- Shared rooms are the least common type of listing, making up only around 0.8% of all listings.

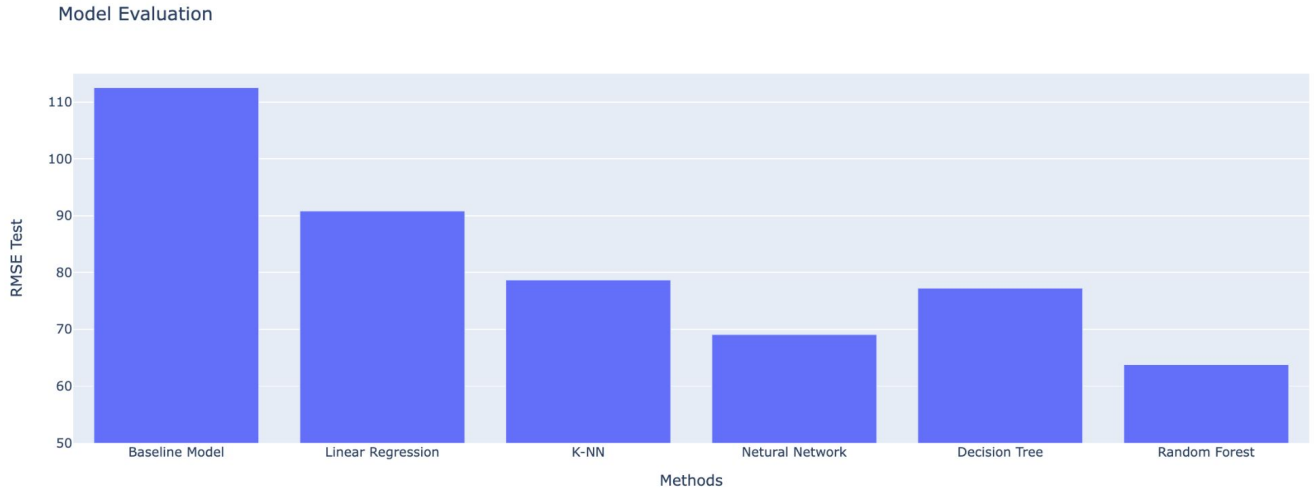


03

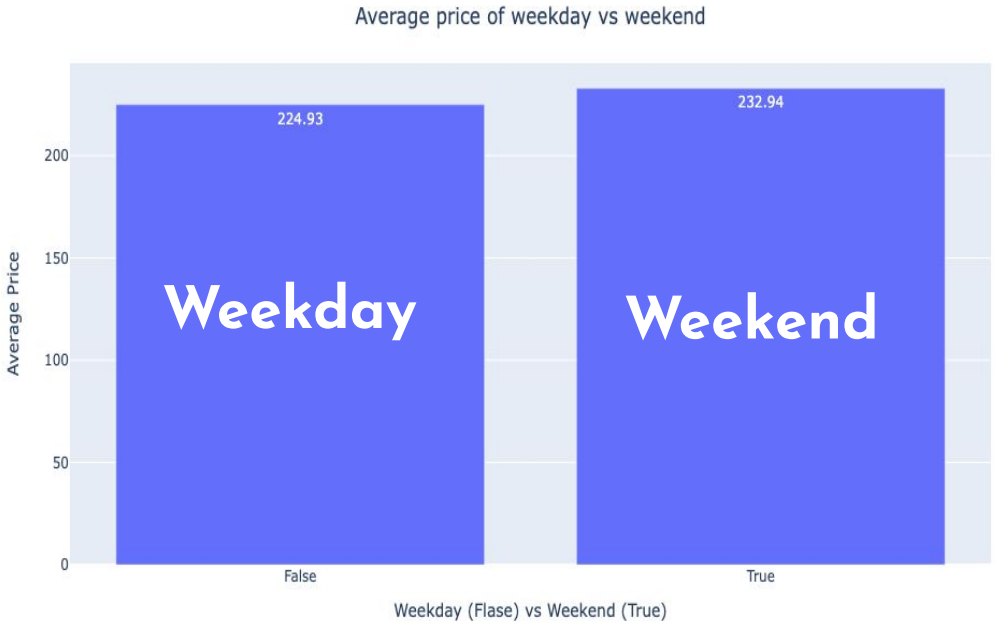
Building Predictive Models

Model Performance Comparison (RMSE)

Model	Baseline	Linear	KNN	Neural Network	Decision Tree	Random Forest
RMSE with PCA	112.5	44.93	37.28	25.98	44.68	31.59
Parameters			K=3	3 hidden layers with 4 nodes	Depth = none	Depth = none



Baseline Model



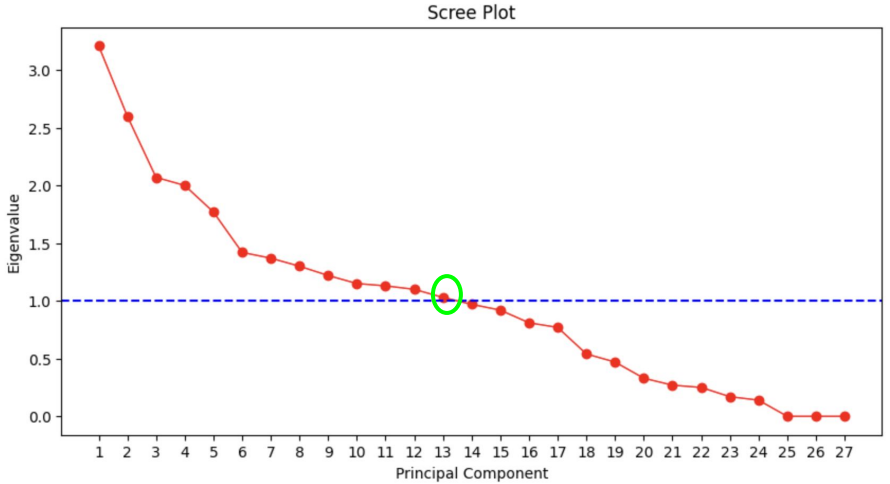
- a DummyRegressor with the strategy of 'mean' is used as a baseline model predicts the mean target value of the training set for all samples in the testing set.
- This is a simple and naive model that can be used as a benchmark to compare the performance of more sophisticated models.

PCA

	% of variance explained	Cumulative % explained		% of variance explained	Cumulative % explained		% of variance explained	Cumulative % explained
0	0.119	0.119	10	0.042	0.712	19	0.012	0.969
1	0.096	0.215	11	0.041	0.753	20	0.010	0.979
2	0.077	0.292	12	0.038	0.791	21	0.009	0.989
3	0.074	0.366	13	0.036	0.827	22	0.006	0.995
4	0.065	0.431	14	0.034	0.861	23	0.005	1.000
5	0.053	0.484	15	0.030	0.891	24	0.000	1.000
6	0.051	0.535	16	0.028	0.919	25	0.000	1.000
7	0.048	0.583	17	0.020	0.939	26	0.000	1.000
8	0.045	0.628	18	0.018	0.957			

- The first twenty-two principal components together explain 99.5% of the variance.
- We utilized the latent root criterion to select the first twelve principal components with eigenvalues larger than 1.

We choose to use PCA.



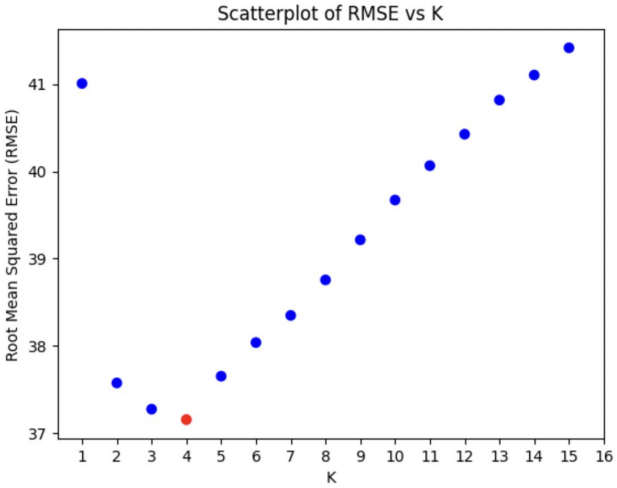
Linear Regression Comparison Before and After PCA

Model MSE: 112.5444

Model MSE: 44.9305

KNN

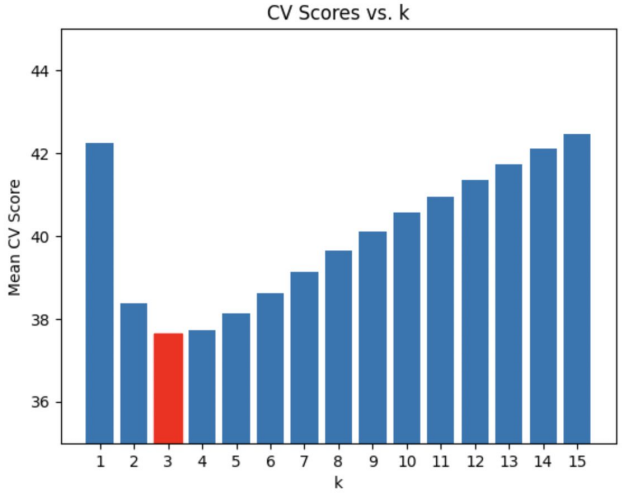
K	RMSE	K	RMSE
1	41.007	9	39.215
2	37.576	10	39.671
3	37.276	11	40.065
4	37.156	12	40.426
5	37.653	13	40.817
6	38.039	14	41.103
7	38.350	15	41.415
8	38.756		



- The root mean square error (RMSE) shows a significant decrease from K = 1 to K = 4, indicating an improvement in the accuracy of the model.
- However, as we increase K beyond 4, the RMSE begins to rise gradually, suggesting a decline in the model's performance.

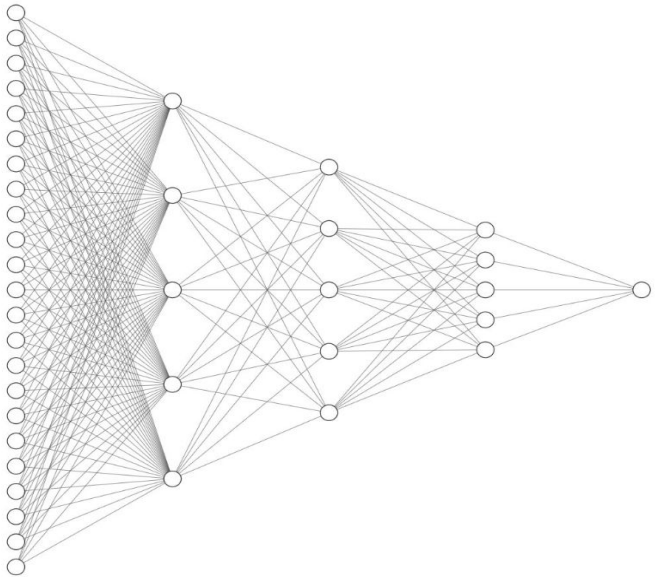
Overfitting

Although K = 4 has the lowest RMSE in KNN performance, consider overfitting and according to the average CV score, K = 3 has the lowest average CV RMSE score which is more accurate. Therefore, we ultimately choose k=3.



Neural Network: 3 hidden layers with 4 nodes returned the lowest RMSE of 25.980.

- Activation: 'logistic'
- Slover: 'adam'
- RMSE decreased with respective to increase the number of hidden layers.
- 3 number of hidden layers with 4 nodes reach the lowest RMSE.

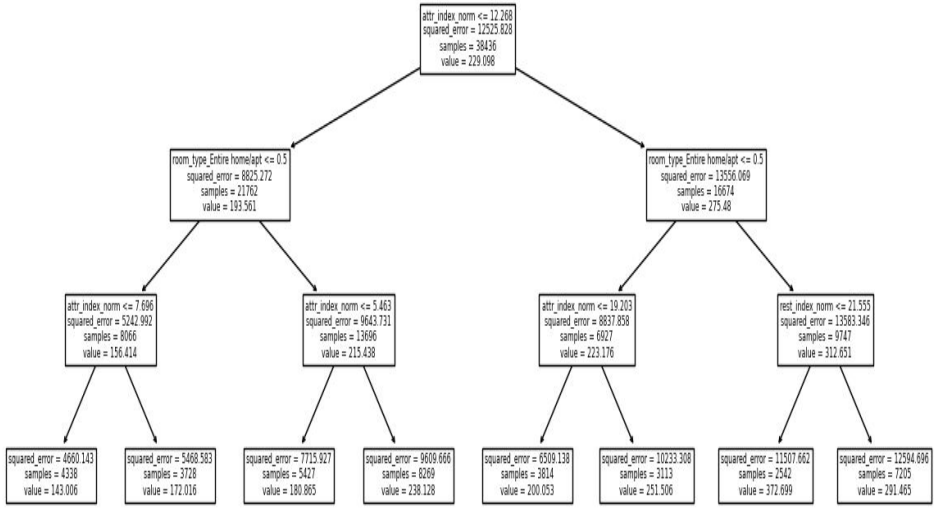


# of hidden layers	# of nodes	RMSE
1	2	44.191
	3	43.583
	4	45.217
	5	41.335
2	2	41.892
	3	32.254
	4	28.680
	5	27.339
3	2	38.387
	3	32.594
	4	25.980
	5	33.188
...	...	...

Decision Tree

GridSearch

The optimal depth is none, with a minimum sample leaf of 10 and a minimum split of 2.



Parameters:	Max. Depth	# of leaves	# of nodes	RMSE
Full Tree	44	34654	69307	68.39
Optimized-Tree	none	3018	6035	**66.17**
Tree with PCA	none	2989	5977	**44.68**

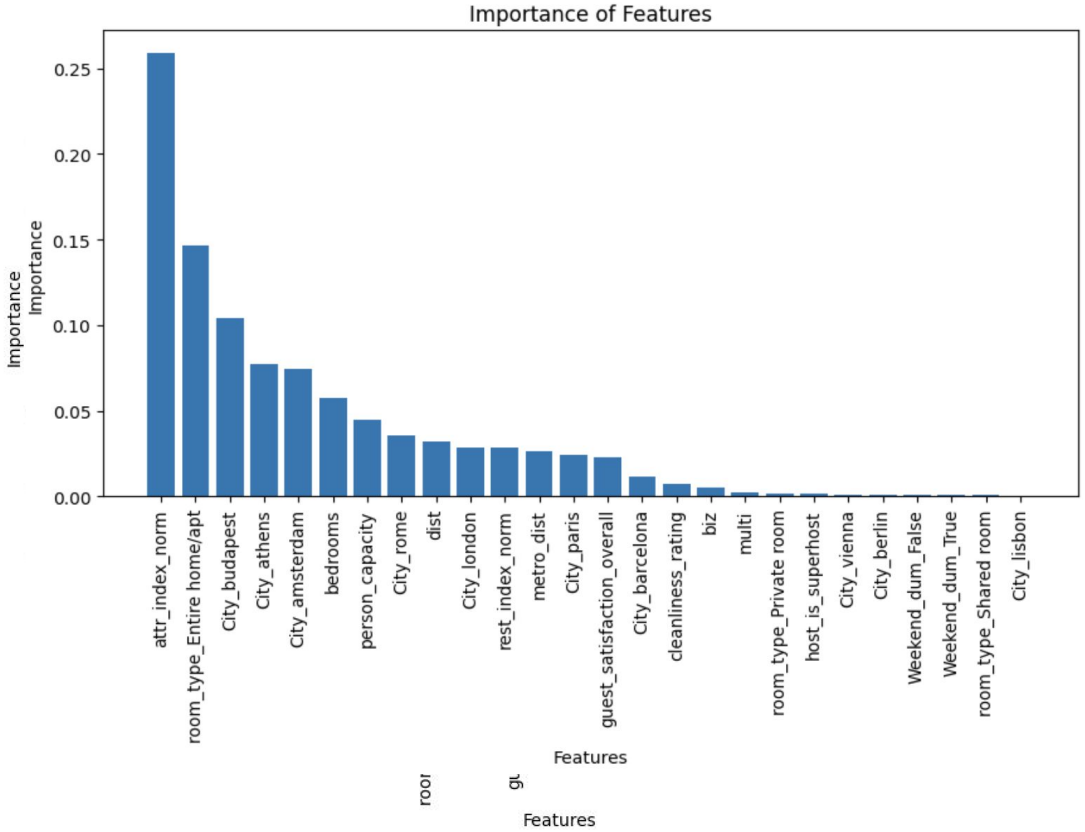
Random Forest

GridSearch with Random Forest

The best parameters for the Random Forest Regressor were determined through a grid search and are as follows: max_depth: **11** min_samples_leaf: **2** min_samples_split: **5** n_estimators: **100**

Parameters: (RMSE)	RMSE
Random forest	64.55
PCA Random forest	31.59

Importance of Features - Random Forest



Top 4 Importance of Features(Excluding cities)

- The quality and quantity of nearby attractions and facilities
- The types of rooms offered, such as the whole house or a single room
- The quantity or number of bedrooms
- The capacity of the unit or quantity of people in the unit can hold

Implications, Limitations and Future Avenues

Implications

- Recognize the factors that affect Airbnb prices in European cities.
- Estimate the price of Airbnb rentals in various cities.

Limitations

- Not be representative of all European cities or all types of Airbnb rentals.
- Not account for all relevant factors that influence Airbnb prices, such as local events, seasonality.
- Suffer from bias, such as inaccurate data or the selection of variables.

Future Avenues

- Expand the dataset to include more cities and a wider range of variables, such as amenities, reviews, and property characteristics.
- Applied to other housing market or different business to predict price.

Potential Improvement

Add specific time of each listing, so we can know the price fluctuation between different seasons and holidays.

Increase more relevant variables.

Regularly update the dataset to enhance accuracy and relevance.

Thank you, any questions?